Data Processing for Big Data Applications using Hadoop Framework

Article · March 2015

DOI: 10.17148/IJARCCE.2015.4343

Navya Francis
KMEA Engineering College

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Data Processing for Big Data Applications using Hadoop Framework

Navya Francis[1], Sheena Kurian K[2]

PG Student,Department of Computer Science, KMEA Engineering College, Edathala, Ernakulam City, India [1]

Associate Professor, Department of Computer Science, KMEA Engineering College, Edathala, Ernakulam City, India [2]

Abstract: The big data is the concept of largespectrum of data, which is being created day by day. In recent years handling these datais the biggest challenge. Hadoop is an open source platform which is used effectively to handle the big data applications. The two core concepts of the hadoop are Mapreduce and Hadoop distributed file system (HDFS). HDFS is the storage mechanism and map reduce is the programming language. Results are produced faster than other traditional database operations. Pig and Hive are the two language which helps us to program the mapreduce framework within short period of time.

Keywords: MapReduce, Pig, Hive, Big data, Hadoop, HDFS.

I. INTRODUCTION

The bigdata [1] contains large spectrum of data, both the structured and the unstructured data. Structured data consist of data in the text and table format. Due to this it can be easily ordered and processed using the data mining tools. Unstructured data does not have an identifiable internal structure, so the processing of these data with traditional databases are not possible.

Data processing is the biggest challenge in bigdata because it contains both types of data, and computations cannot be performed by the customary database and data mining techniques. Research study states that bigdata contents are generated day by day. IBM [2] states that 2.5 billion gigabytes of data are produced in a single day.

Bigdata has several characteristics[3]. Volume refers to the large range of the data produced per second. Variety refers to the different formats of data. For example consider a bank transaction, in this the various forms of transaction are cheque, ATM, paying slip etc. Velocity means the speed of production of data from various machines, sensors, log files etc. Complexity refers to the handling of these huge data.

II.HADOOP FRAMEWORK

Bigdata problems is handled effectively, using the concepts of hadoop. Hadoop [4] is an open source software developed by the Apache. It acts as cross platform operating system. Hadoop contains the distributed file system inorder to handle the large range of data. Hadoop[5] has many features, like reliability, data locality, cost effectiveness and efficient computation etc.

High data locality helps us in fast processing. By simple steps we can process the large data contents and hadoop provides efficient computation of data in highly cost effective manner. Reliable,stable and consistent data is generated, which means data contents will be the same all the time after processing set of inputs.Due to these features hadoop is used to process the bigdata contents.

Hadoop has many different vendors. Cloudera, Horton works, MapR, Amazon Elastic Mapreduce, IBM Infosphere Big Insights are some of them. There are some core components of hadoop, using it we can effectively compute the big data contents in more efficient manner. The two core components of hadoop are, hadoop distributed file system (HDFS) and mapreduce.

III. HADOOP DISTRIBUTED FILE SYSTEM

HDFS was designed in the project NUTCH[6]. It acts as the storage mechanism. Input data are split into different chunks and stored in HDFS[7]. The default chunk size is 64MB. HDFS has block oriented architecture. Each block has fixed size and are stored in the hadoop cluster. These different blocks are called as data node and they contains the actual data. The data nodes are stored in different machines at different clusters. The data is processed in the same cluster were it is stored, due to this it avoid the problems related to transferring of data from one place to another. Thus the HDFS provide reliable and fast access to the stored data.

Name node stores the metadata for the file system across each hadoop cluster. Name node is stored in the main memory, so it allows fast random access. The data stored in the name node are persistent and due to this failure will result in the permanent loss of the data.Because it contains all the links to the data nodes. To avoid the loss of information, the secondary name node is maintained. It contain the image of the name node and the edit logs. When failure comes, based on these log details the data can be retrieved. The secondary name node cannot be replaced directly instead of the actual name node.

HDFS has the master slave architecture[8]. Each hadoopcluster contains a single name node,that is the master node andslave nodes are the data nodes. The primarycommunication mechanism between the name node and data node is called heartbeat. In every three seconds the heartbeat is sent to the name node from the data node. Heart beat contain the block report and list of blocks in the data node. If the heartbeat is not received the name node will create another replica of the data node. The name node will always maintain always three copies

DOI 10.17148/IJARCCE.2015.4343

of the data nodes. Due to this single point failure will not affect the HDFS.

Rack awareness is another important feature of the HDFS. Different copies of data are stored in different racks by HDFS with different rack id and bandwidth. Different racks have different bandwidth. We know that HDFS always keep three copies of data. Instead of saving the data and copies in same rack, we can save them in different racks. Due to this single rack failure will not affect the losing of data. The overhead of saving the data in different racks are avoided.

IV. MAPREDUCE

The mapreduce framework helps us to retrieve the data efficiently from large spectrum of data. The mapreduce framework was introduced by Google in 2004 [9]. Mapreduce process the data as the key value pair. There are mainly four phases to perform the mapreduce operation. First phase, mapper phase will collect the data from the HDFS which are stored in the different clusters. Output from the mapper phase is the intermediate results. These results are then given to other phase for passing to the reducer.

Second phase is shuffle phase, here intermediate results are shuffled so that the results of the different mappers are brought together. Third phase is sort phase. In this the shuffled intermediate results are sorted together based on the key value, so the same key valued contents are brought together. By doing sorting the contents can be easily passed to the reducer for processing.

Last phase is reducer phase. In this the sorted contents are processed to get the significant data. The jobs in the hadoop cluster are performed by the task tracker. When a job is scheduled, the job tracker will assign the job to the task tracker. It will proceedsto the job execution and the output is produced. Thus the large range of data is processed into useful contents.

The algorithm was written in JAVA, but it require lot of time to create the mapreduce framework. Framework which was written are difficult to understand and it requires a lot of time for execution. Inorder to overcome these problems the mapreduce framework are effectively implemented using the Hive and Apache Pig. These two languages helps us to easily retrieve the data from the HDFS using the mapreduce framework

A.HIVE

Hive is used to retrieve the big data contents from the hadoop cluster. Hive provides the SQL dialect called hive query language (HQL) [10], which retrieve the data. It helps us to move the existing database into the hadoop without much variations. This is most suited for the data warehouse applications, where large datasets are considered and the static data are analysed.

Hive has certain limitation. Basic operations like insert, delete, update options are not available. It has higher latency and does not provide transactions. Hive has primitive and collective datatypes like the SQL language. Primitive datatypes support the integer, floating point, boolean type and character strings of arbitrary. Timestamps and binary fields are added in the latest version of hive.

Collective datatypes include struct, array and map**. By** using these datatypes we can break the normal form of the traditional database. Due to this data duplication takes place and it consumes disk storage and potential data inconsistency. Hive database concept is namespace of tables, due to this collision of the table names are avoided. Thus different users of multiple teams in a large cluster can use the database effectively. If the database is not specified it uses the default database.

Hive have options to alter the database, create tables, alter the table, drop the table, partition and manage the tables. Hive has some additional features like we can insert data into table through queries, exporting data are some of them.

B.PIG

Apache Pig is open source [11]. It provides an engine for hadoop for parallel execution of data flows. The scripting language used in pig is the pig latin [12]. It is very simple programming language and the developer can easily code the mapreduce framework in simple steps. Due to this the code can be easily understood and the corrections can be cleared faster. These are some of the advantages of the pig. Since the pig is easy to understand than other programming language it is widely used inorder to perform the mapreduce functions. It has several additional functions[13] like join, filter, cogroup, orderby etc. to perform the operations of the mapreduce efficiently.

Join function helps us to combine more than one files together, based on the requirement. There are different join functions like outer join, inner join etc. By using these join functions we can effectively combine different files. Filter function will remove the corrupted or incomplete records from the data considered for processing. Order by will sort the contents retrieved in the order which we want to process. Cogroup is another important feature of pig. In this two different groups are combined together to perform the execution.

Pig has many build in user defined functions and piggybank etc. Built in functions to load input data, store output data, filter only the suitable contents and evaluation of these contents before processing. Piggybank contains the user contribution functions. It is a part of pig distribution and they are not built in functions. Inorder to access them we have to load the piggybank.jar to the registry.

V. MODELLING OF HADOOP FRAMEWORK

The hadoop framework is used to process the big data applications. It joins multiple datasets together. There are different step in the modelling of the hadoop framework. First is the storing of the contents into the HDFS. After the contents are stored, we can process the data using the mapreduce concept. HDFS splits the contents into different chunks and save in different data nodes of the hadoop cluster.

DOI 10.17148/IJARCCE.2015.4343

Second is the mapreduce algorithm. This will map the contents from different data nodes as the key value pair and reducer will process the contents to get the meaningful data. The block diagram of the hadoop framework is shown in Figure.1

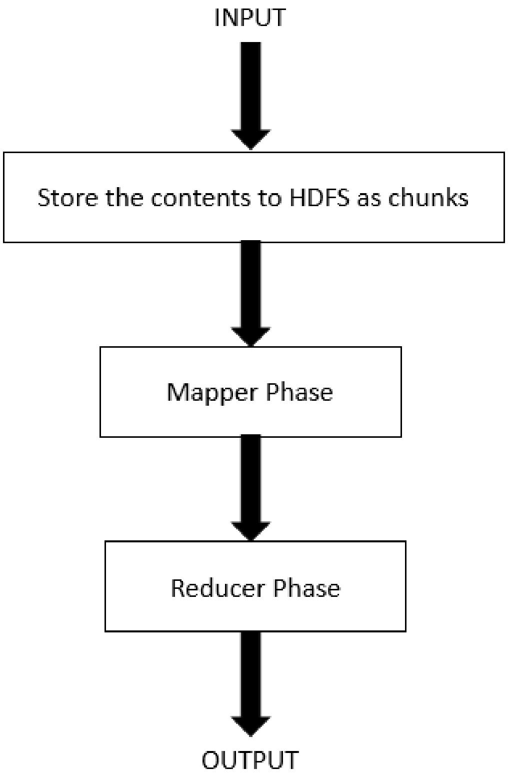

Figure.1 Block Diagram of modelling hadoop framework

During the execution, both file and HDFS perform read and write operation. The parameters received by the simple mapreduce execution is shown in the Table.1

TABLE 1
Mapreduce Execution Parameters

Parameters Received	Map (Bytes)	Reduce (Bytes)
File Bytes Read	0	1275255
File Bytes Written	1559129	1558808
HDFS Bytes Read	12120698	0
HDFS Bytes Written	0	812789

VI. CONCLUSION

Hadoop is the tool used to manage and process the bigdata contents, which is the biggest challenge in the recent years. By using Hadoop distributed file system and map reduce concepts in hadoop we can process any bigdata contents within short period of time. HDFS acts as the storage mechanism in the hadoop and mapreduce is used as the programming language inorder to process the contents. Mapreduce is operated with the help of two functions, mapper function and the reducer function.

VII. FUTURE WORKS

Hadoop is widely used for strategic decision making in the big data applications.It has many application areas like fraud detection, pattern recognition, content optimizing, marketing analysis, network analysis, large data transformations etc. are some of them.

Hadoop framework can be used to make informed decisions in logistic freight inorder to perform the freight audit. By doing freight audit they can prevent organizations from overpaying for the services of freight forwarders, which they haven't used.

REFERENCES

[1] http://en.wikipedia.org/wiki/Big_data Retrieved 2015-03-12.
[2] http://www.bbc.com/news/business-26383058 Retrieved 2015-03-12.
[3] Shital Suryawanshi, Prof.V.S.Wadne "Big Data Mining using Map Reduce: A Survey Paper" IOSR International Journal Computer Engineering, Volume 16 Issue 06, Nov- Dec 2014, Pages 37-40.
[4] http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html Retrieved 2015-03-12.
[5] http://en.wikipedia.org/wiki/Apache_Hadoop Retrieved 2015-03-12.
[6] http://wiki.apache.org/nutch/NutchTutorial Retrieved 2015-03-12.
[7] Ramesh Kumar, Dr.Vijay Singh Rathore "Efficient Capabilities of Processing of Big data using Hadoop Map Reduce" International Journal of Advanced Research in Computer and Communication Engineering, Volume 03 Issue 06, June 2014, Pages 7123-7126.
[8] Boris Lublinsky, Kevin T Smith, Alexey Yakubovich "Professional Hadoop Solutions" Proc.WROX, Pages 1-96.
[9] http://en.wikipedia.org/wiki/MapReduce Retrieved 2015-02-27.
[10] Edward Capriolo, Dean Wampler, Jason Rutherglen "Programming Hive" O'Reilly Media, Edition 1, October 2012.
[11] http://en.wikipedia.org/wiki/Pig (programming_tool) Retrieved 2015-02-27.
[12] http://pig.apache.org/docs/r0.8.1/udf.html Retrieved 2015-02-26.
[13] Alan Gates "Programming Pig" Proc. O'Reilly Media. Pages 1-170.
[14] Yaxiong Zhao, Jie Wu "Dache: A Data Aware Caching for Big-Data Applications Using The Map Reduce Framework" International Journal of Tsinghua Science And Technology, Volume 19 Number 1, February 2014, Pages 39-49.
[15] Jefffrey Dean, Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters" Communications of the ACM, Volume 51, Number 1, Pages 107-113.
[16] Sasiniveda.G, Revathi.N "Data Analysis using Mapper and Reducer with Optimal Configuration in Hadoop" International Journal of Computer Trends and Technology (IJCTT), Volume 04 Number 03, February 2013, Pages 264-268.
[17] Karan B.Maniar, Chintan B.Khatri "Data Science: Bigtable, Mapreduce and Google File System" International Journal of Computer Trends and Technology (IJCTT), Volume 16 Number 03, October 2014, Pages 115-118.
[18] Tom White "Hadoop the definitive guide" Proc. O'Reilly Media, Edition 3, May 2012.
[19] Chuck Lam "Hadoop in Action" Proc. Manning Publication, Edition 1, December 2012.
[20] Donald Miner, Adam Shook "Mapreduce Design Patterns" Proc. O'Reilly Media, November 2012.
[21] http://en.wikipedia.org/wiki/Logistics Retrieved 2015-02-27.
[22] http://en.wikipedia.org/wiki/Freight_rate Retrieved 2015-02-27.

DOI 10.17148/IJARCCE.2015.4343

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

[23] http://hortonworks.com/hadoop-tutorial/how-to-use-basic-pig-commands/ Retrieved 2015-02-26.

[24] https://www.controlpay.com/services/logistics-visibility Retrieved 2015-02-26.

[25] http://en.wikipedia.org/wiki/Big_data Retrieved 2015-02-26.

[26] http://www-01.ibm.com/software/in/data/bigdata/ Retrieved 2015-02-26.

[27] http://bigdatauniversity.com/bdu-wp/bdu-course/big-data-fundamentals/ Retrieved 2015-02-26.

[28] Sangeeta Bansal, Dr. Ajay Rana “Transitioning from Relational Databases to Big Data” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014.

[29] Han Hu, Yonggang Wen, Xuelong Li “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial” IEEE access practical innovation: open solution, Volume 2, July 2014, Pages 652-687.

[30] Kyongha, yoonjoon, Hyunsik, Yondohn, Bongki “Parallel Data Processing with MapReduce: A Survey” ACM SIGMOD Record, Volume 40 Issue 4, December 2011 Pages 11-20.

DOI 10.17148/IJARCCE.2015.4343